

A MAXIMUM LIKELIHOOD METHOD FOR ANALYZING PSEUDOGENE EVOLUTION: IMPLICATIONS FOR SILENT SITE EVOLUTION IN HUMANS AND RODENTS

BU- 1569-M

May 2001

**Carlos Bustamante
Rasmus Nielsen
and
Daniel L. Hartl**

Keywords: Pseudogene evolution, synonymous site evolution, codon based models, maximum likelihood, neutral evolution

Abstract:

We present a new likelihood method for detecting constrained evolution at synonymous sites and non-neutral evolution in putative pseudogenes. The model is applicable when DNA sequence is available from a protein-coding functional gene, a pseudogene derived from the protein-coding gene, and an orthologous functional copy of the gene. Two likelihood ratio tests are developed to test the hypotheses that the putative pseudogene evolves neutrally and that the rate of synonymous substitution in the functional gene equals the rate of substitution in the pseudogene. The method is applied to a data set containing 74 human processed pseudogene loci, 25 mouse processed pseudogene loci and 22 rat processed pseudogene loci. Using a working draft of the human genome, we localized 67 of the human/pseudogene pairs and estimated the GC content of the surrounding genomic region. We find that for pseudogenes that land in GC regions similar to those of their paralogs, the assumption of equal rates of silent and replacement site evolution in the pseudogene are upheld and that the rate of silent site evolution in the functional genes is ~70% the rate of the pseudogene. For pseudogenes that land in genomic regions of much lower GC than their functional gene, we see a sharp increase in the rate of silent site substitutions leading to a large rejection rate for the pseudogene neutrality test.

**A Maximum Likelihood Method for Analyzing Pseudogene Evolution: Implications for
silent site evolution in humans and rodents**

Pseudogene.pdf

Carlos D. Bustamante¹, Rasmus Nielsen², and Daniel L. Hartl¹

¹*Department of Organismic and Evolutionary Biology, Harvard University and*

²*Department of Biometrics, Cornell University.*

Running Head: Rate of Synonymous Substitution

Keywords: Pseudogene evolution, synonymous site evolution, codon based models, maximum likelihood, neutral evolution

Corresponding author:

Daniel L. Hartl

16 Divinity Ave.

Cambridge, MA 02138, USA

Fax: 617-496-5854

E-mail: dhartl@oeb.harvard.edu

Abstract

We present a new likelihood method for detecting constrained evolution at synonymous sites and non-neutral evolution in putative pseudogenes. The model is applicable when DNA sequence is available from a protein-coding functional gene, a pseudogene derived from the protein-coding gene, and an orthologous functional copy of the gene. Two likelihood ratio tests are developed to test the hypotheses that the putative pseudogene evolves neutrally and that the rate of synonymous substitution in the functional gene equals the rate of substitution in the pseudogene. The method is applied to a data set containing 74 human processed pseudogene loci, 25 mouse processed pseudogene loci and 22 rat processed pseudogene loci. Using a working draft of the human genome, we localized 67 of the human/pseudogene pairs and estimated the GC content of the surrounding genomic region. We find that for pseudogenes that land in GC regions similar to those of their paralogs, the assumption of equal rates of silent and replacement site evolution in the pseudogene are upheld and that the rate of silent site evolution in the functional genes is ~70% the rate of the pseudogene. For pseudogenes that land in genomic regions of much lower GC than their functional gene, we see a sharp increase in the rate of silent site substitutions leading to a large rejection rate for the pseudogene neutrality test.

Introduction

It has long been held that pseudogenes provide the molecular evolutionist with an important tool for studying the rate and pattern of neutral evolution (Li, Gojobori, and Nei, 1981). To the extent that pseudogenes evolve without selective constraint, the rate and pattern of substitutions in pseudogenes will faithfully reflect the underlying mutational process.

Pseudogenes have, therefore, been used to infer the mutational process for nucleotide changes within species (Gojobori, Li, and Graur, 1982; Li, Wu, and Luo, 1984) and to compare this process between species (Petrov and Hartl, 1999) as well as to study deletion rates among taxa (Graur et al., 1989; Ophir and Graur, 1997; Petrov, Lozovskaya, and Hartl, 1996) and the effect that rates of DNA loss have on genome size (Petrov et al., 2000; Bensasson, 2001).

Pseudogenes also provide the molecular evolutionist with a direct opportunity to infer the strength of selection on changes at synonymous sites. Ophir et al. (1999), employing a distance method to analyze a set of 12 human and murid (rat and mouse) pseudogene gene trees, found that, on average, murid and human third-position sites evolve at 40% the rate of pseudogene third position sites. Since not all changes at third-position sites are synonymous, it is difficult to extrapolate from this result how much selection is acting on synonymous changes.

This question is of considerable practical importance in the study of molecular evolution, since the ratio (ω) of the number of replacement substitutions per replacement site (d_N) to number of synonymous substitutions per synonymous site (d_S) is useful in detecting adaptive evolution at the protein level. If there is considerable variation in selection intensity at synonymous sites across genes, this approach is compromised, and it becomes quite difficult and perhaps meaningless to compare ω between genes or species.

The issue of whether silent sites are evolving at the neutral mutation rate is far from resolved. The analysis of polymorphism data and codon frequencies in *E. coli* and *Salmonella enterica* genes suggest that there is considerable weak selection operating on silent sites (Andersson and Kurland, 1990; Sharp et al., 1993; Hartl, Moriyama, and Sawyer, 1994; Hart et al., 2000). Recent work in comparative analysis of *Drosophila* genes has also revealed evidence for constraint on synonymous site evolution, presumably due to codon bias (Akashi, 1996; Akashi, 1997; Akashi and Schaeffer, 1997; McVean and Vieira, 2001). In rodent genes, however, there seems to be no relationship between the rate of synonymous substitution and codon bias (Smith and Hurst, 1999). It has been suggested that a mutation/selection equilibrium may account for genomic variation in GC content and consequently affect the rate of substitution at silent sites (for a review see Bernardi, 2000).

Current methods used to study pseudogene evolution have not exploited recent statistical developments in molecular phylogenetics. In this paper, we develop codon-based models for gene-trees that contain a (processed) pseudogene, the functional gene from which the pseudogene was derived, and the ortholog of the functional gene from a closely related species (Fig. 1). The models are implemented in a maximum likelihood framework and lead to two likelihood ratio tests of neutral evolution—one to test if the pseudogene is evolving neutrally and one to test if the synonymous sites in the functional gene are evolving at the same rate as the pseudogene. We apply this method to a dataset consisting of 121 processed pseudogenes: 74 from human, 22 from rat, and 25 from mouse (Ophir and Graur, 1997) to test the assumption that processed pseudogenes evolve neutrally, and to estimate the strength of selection on synonymous sites in the functional paralogs of pseudogenes. Using a draft of the human genome, we were able to localize 67 of the gene / pseudogene pairs and estimate the GC content of the surrounding

genomic areas for both. We used this data as well as the GC content of fourfold redundant sites in the functional genes to assess whether rejection of pseudogene or silent site neutrality is correlated with GC content.

Statistical Methods

The Model

The method we will use to analyze pseudogene evolution is based on the likelihood models developed by Goldman and Yang (1994), Muse and Gaut (1994), Nielsen and Yang (1998), Yang (1998), and Yang and Nielsen (1998). In these models, the DNA sequence is treated as a sequence of triplets of nucleotides (codons). We will assume that the substitution processes in each codon site are independent and can be described by a continuous time Markov chain with state space on the 61 codons of the standard genetic code (excluding stop-codons). Furthermore, we assume that the process can be parameterized in terms of the transition/transversion rate ratio (κ), the d_N/d_S ratio (ω) and the stationary frequencies of each codon (π_i). The transition rate matrix $Q = \{q_{ij}\}$ is then defined as

$$q_{ij} = \begin{cases} 0, & \text{if codons } i \text{ and } j \text{ differ at more than one codon position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases} \quad (1)$$

The transition probabilities of the process can then be calculated by exponentiating the rate matrix, using standard numerical methods (e.g., numerical diagonalization of Q). The codon frequencies are estimated from the data using method of moments to reduce the number of parameters in the model and to save computational time. The estimates are obtained from the observed base frequencies in the three codon positions (see Yang and Nielsen 1998 for details).

For the purpose of analyzing pseudogenes, we assume a phylogenetic tree in which there is a pseudogene, a functional paralog from the same species, and an ortholog to the functional gene from a related species (outgroup) (Fig. 1). In the most general model we will assume that the values of ω and the branch length of the three branches (t_f , t_o and t_ψ) are independent parameters, and that the codon frequencies and κ do not vary among branches. The set of parameters that will be estimated by maximum likelihood is then $\Theta = \{t_f, t_o, t_\psi, \omega_f, \omega_o, \omega_\psi, \kappa\}$. The likelihood function to be optimized is proportional to

$$\Pr(S | \Theta) = \prod_{j=1}^n \sum_{i=1}^{61} \prod_{k=1}^3 \pi_i P_\Theta(c_i \rightarrow S_{kj}), \quad (2)$$

where n is the number of codons in each sequence, S is the $3 \times n$ matrix of sequence data containing the codon in sequence k at position j in entry S_{kj} and $P_\Theta(c_i \rightarrow S_{kj})$ is the transition probability along the appropriate branch of the phylogeny from codon c_i ($i = 1, 2, \dots, 61$) in the internal node to codon S_{kj} . Parameter estimates are obtained by maximizing the logarithm of Equation (2) with respect to Θ .

Likelihood Ratio Tests

Assuming the putative pseudogene is truly a pseudogene, it is reasonable to assume $\omega_\psi = 1$, i.e., the rate of synonymous substitution is equal to the rate of replacement substitution. If the gene is not transcribed, there should be no selection acting on the protein level. Furthermore, if the pseudogene arose after the two species diverged, the pseudogene sequence and the functional ortholog are more closely related to each other than either of them is to the outgroup sequence. If the rate of evolution of the pseudogene equals the rate of synonymous evolution in

the functional gene, then $t_f = t_\psi$. Equality of t_f and t_ψ is the major hypotheses we will test in this paper. In this model, branch lengths are not scaled by the expected number of substitutions. For a particular lineage in the phylogeny, the maximum likelihood estimate of the number of synonymous substitutions per synonymous sites (d_S) and the number of replacement substitutions per replacement sites (d_N), can then be calculated from the maximum likelihood estimates of t , κ and ω using the methods described in Yang and Nielsen (1998).

We will compare the maximum likelihood under a constrained model with 6 parameters $\{t_f, t_o, t_\psi, \omega_f, \omega_o, \kappa\}$ to the maximum likelihood under a nested model with 5 parameters $\{t_f = t_\psi, t_o, \omega_f, \omega_o, \kappa\}$. Appealing to the usual large sample results for nested hypotheses (e.g., Stewart, Ord, and Arnold, 1999, pp. 246), two times the logarithm of the maximum likelihood ratio of the two hypotheses (LRT[5,6]) is asymptotically distributed as a χ^2 random variable with one degree of freedom (d.f.). In particular, if the maximum log likelihood ratio under the 6 parameter model is more than 1.92 (3.84/2) log likelihood units larger than the maximum likelihood value under the 5 parameter model, we reject the null hypotheses ($t_f = t_\psi$) at the 5% significance level. LRT[5,6] is a test of the hypothesis that the rate of substitution is the same in synonymous sites and pseudogene sites. Throughout the rest of this paper, we will refer to LRT[5,6] as the silent site neutrality likelihood ratio test (SSNLRT).

A second hypothesis we will test is that the pseudogene has been an untranscribed pseudogene or a neutrally evolving gene in the entire evolution of the pseudogene lineage. To do this we compare the log maximum likelihood of the general model with 7 parameters ($t_f, t_o, t_\psi, \omega_f, \omega_o, \omega_\psi, \kappa$) to the log maximum likelihood of the nested 6 parameter model ($t_f, t_o, t_\psi, \omega_f, \omega_o, \kappa$) which assumes $\omega_\psi = 1$. Again, significance is tested by comparing twice the log likelihood ratio, LRT[6,7], to a χ^2_1 -distribution. If the assumptions of the model are correct and we reject the

null hypothesis ($\omega_\psi = 1$), then we would conclude that selection must have been acting at the protein level in the lineage leading to the sampled pseudogene. Throughout the rest of the paper, we will refer to LRT[6,7] as the pseudogene neutrality likelihood ratio test (PNLRT).

Combined Data Analysis

One issue of interest is whether we can reject the null hypotheses when combining the data across loci. If we assume that the genes are independent of one another, we can sum the log-likelihoods under each model and perform the silent site neutrality (SSNLRT) and the pseudogene neutrality (PNLRT) likelihood ratio tests on the pooled data. SSNLRT applied to a combined dataset would test whether all pseudogenes in the dataset conform have $\omega_\psi = 1$. PNLRT applied to the whole dataset would test whether the average rate of silent site evolution in each gene is the same rate as in the respective pseudogene for all genes. Since models 6 and 7 differ by one degree of freedom, the PNLRT for n genes will be distributed as χ_n^2 . Likewise, the combined SSNLRT statistic will be χ_{n-r}^2 distributed where r is the number of genes that reject the PNLRT, since we would not perform the SSNLRT on a gene if the gene rejects the pseudogene neutrality likelihood ratio test.

Quantile-quantile Plots for the ecdf and cdf χ_1^2

A quantile-quantile plot (q-q plot) is a way of comparing the cumulative distribution function for two random variables. If the two random variables have the same distribution, their q-q plot lies on the diagonal line $x = y$ in standard Cartesian coordinates. The empirical cumulative distribution function (ecdf) of a sample is the analog of the cumulative distribution function for a random variable. We would therefore expect that if the all genes conform very well to the null hypothesis for a particular test we would expect that a q-q plot of the ecdf for the distribution of the LR statistic across genes vs. the cdf of a χ_1^2 -distribution to lie on the diagonal

line. In Fig. 3, we present such q-q plots for the distribution of LRT[5,6] and LRT[6,7] among genes as compared to χ^2_1 . As outlined in Rice (1995), the ecdf for a sample of numbers x_1, x_2, \dots, x_n is defined as:

$$F_n(x) = \frac{1}{n} (\# x_i \leq x) \quad (3)$$

Letting $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ represent the order statistics (ranks) of the x_i 's, then for any $x_{(k)} \leq x < x_{(k+1)}$, $F_n(x) = k/n$.

Estimating selection on synonymous sites

A natural estimator of the average strength of selection on synonymous sites is the ratio of the average number of substitution per synonymous site in functional genes to the average number of substitution per site in the pseudogenes across gene pairs ($\overline{ds_f} / \overline{ds_\psi}$). The reason we use the ratio of the averages instead of the average of the ratios is that for most distributions the former statistic should converge in large samples to the “true” ratio of the means whereas the latter statistic will not necessarily converge. To generate confidence intervals for this statistic, we used 100,000 non-parametric bootstrap samples generated by sampling with replacement (ds_f, ds_ψ) pairs estimated from our data.

Material and Methods

Data

For each pseudogene reported in Ophir and Graur (1997) we searched the non-redundant database (nr) of the NCBI server (<http://www.ncbi.nlm.nih.gov/BLAST/>) using the program `blastn` (Altschul *et al.* 1990) to find the closest available paralogous sequence from the same species. We then ran `blastn` on the paralog to find the single closest murid or human ortholog

for the paralog/pseudogene pair (Genbank accession numbers available from the authors upon request). Genes for which no functional ortholog could be found—as evidenced by the intercalation of phylogenetically incongruent sequences in the search results—were omitted (e.g., if a search using a functional human gene revealed a higher homology between the human gene and a crocodile gene rather than the human gene and the mouse gene, the pseudogene/ortholog pair was omitted).

Sequences were obtained, edited, and aligned by CLUSTALW using various versions (3.5-4.0.30) of the program DAMBE (Xia, 2000). All stop codons and codons containing nucleotides with gaps in the alignment were removed and the reading frame set to the reading frame of the mouse or rat and human functional gene. Pseudogenes that were identical to their functional paralog after gaps were removed from the alignment were omitted from the analysis as were genes for which the estimated pseudogene t was longer than the outgroup t in Model 7. We used `blastn` to identify the position of 67 of the human pseudogene/functional gene pairs using various working drafts of the human genome. All genomic GC content measures are for the NCBI Genbank entry for the BAC (> 100,000 bp) containing the functional gene.

Results and Discussion

In Appendix 1, we summarize maximum likelihood parameter estimates and the results of the individual likelihood-ratio tests for each gene. Overall, the pseudogene neutrality likelihood ratio test (PNLRT) is significant at the 5% significance level for 13 of the 121 genes (11%). For those genes for which the pseudogene neutrality likelihood ratio test is not significant ($n = 108$), 25 genes (23%) have significant silent-site neutrality likelihood ratio tests (SSLRT). Given that we have performed a large number of tests, we would expect approximately 5% of the genes to

reject either null hypothesis if the null hypotheses were true simply due to chance. Ninety-five percent confidence intervals for the “true” proportion of genes that reject the null hypothesis, though, do not cover 5% in either test, suggesting that we have rejected the null hypotheses in an unusually large number of genes for both tests.

Implications of the results for Pseudogene Evolution

To see if GC content correlates with an increase in the rejection of either test, we divided the dataset into two equal groups based on the GC content of fourfold redundant sites in the functional paralogs of the pseudogenes (GC4). The rationale for using GC4 rather than overall GC content is that replacement sites are presumably constrained by purifying selection on amino acid changes. For 67 of the human genes we were also able to estimate the GC content of the genomic region for the functional gene using a working draft of the human genome. Not surprisingly, GC4 was highly correlated with the genomic GC content ($r = 0.72$) such that the results that follow are qualitatively similar and the conclusions unaltered if one uses flanking GC content of the functional gene rather than GC4.

In Table 1 we report the results of the pooled likelihood ratio tests for gene pairs grouped by GC content of the functional gene. We find that regardless of whether we consider the regional GC content or GC4 of the functional gene in humans, processed pseudogenes derived from GC-rich genes tend to reject the pseudogene neutrality ratio test more frequently than pseudogenes derived from GC-poor genes. A pooled likelihood ratio test suggests that pseudogenes derived from GC-rich genes, in general, do not have equal rates of substitution at silent and replacement sites as one would expect under a model of no selection at the amino acid level ($p < 0.0001$). The same phenomenon is also observed in the rodent pseudogenes ($p < 0.0032$).

This pattern is not observed in genes derived from GC-poor regions. In fact, the low level of rejection of the pseudogene neutrality likelihood ratio test for genes derived from GC-poor regions is not significantly different from what would be expected by chance. Likewise, pooled pseudogene neutrality likelihood ratio tests suggest that pseudogenes derived from GC-poor regions conform to the expectation of equal rates of silent and replacement substitution. An important implication of this result is that one should be very cautious in using pseudogene data to model neutral evolution, particularly when one cannot test whether the pseudogenes conform to the expectations of neutral theory.

At least two possible hypotheses explain the observed excess of non-neutral pseudogenes (i.e., pseudogenes that have $\omega \neq 1$ and reject PNLRT). The first explanation is that the assumed phylogeny for some pseudogenes is incorrect (Fig. 3). One reason the phylogeny may be incorrect is that the duplication or introgression event that gave rise to pseudogene may predate speciation. In this case, the two orthologous functional genes would be more closely related to each other than either is to the pseudogene and the assumptions for Model 6 would not be met. Likewise, the pseudogene may be derived from another closely related functional gene rather than from the paralog used in the analysis. If the gene duplication event that gave rise to the two paralogous functional genes predates speciation, we would again have a scenario where the two orthologous functional genes are more closely related to each other than either is to the pseudogene. If the gene duplication event occurred after speciation, the inferred pseudogene branch would contain the shared history between the duplication event that gave rise to the functional gene and the duplication event that gave rise to the pseudogene, confounding either adaptive or constrained evolution and neutral evolution. Two predictions of this hypothesis are (1) that the majority of processed pseudogenes that reject the pseudogene neutrality likelihood

ratio test will have ω less than 1 in the pseudogene branch, and (2) that the t_ψ parameter (length of the time the process has been occurring along the pseudogene branch) in full model will be longer than the t_o parameter.

In Figure 5, we plot the fraction of genes that reject Model 6 for a given ω range. As predicted, in every instance that model 6 was rejected, ω was less than 1. Unfortunately this prediction is also consistent with the second hypothesis we will discuss shortly. The second prediction of the hypothesis is not supported. In fact, only five of pseudogenes we have analyzed had $t_\psi > t_o$, and those genes were not included in the likelihood ratio tests presented in Table 1 suggesting that the first hypothesis does not explain the excess number of pseudogenes that reject the PNLRT. This hypothesis is, also, not consistent with the observed result that pseudogenes from GC poor regions do not reject the pseudogene neutrality likelihood ratio while genes from GC rich regions do.

This does not mean, though, that the problem is non-existent altogether. We did find a clear example of an incorrectly inferred phylogeny in the Calmodulin 2 pseudogene (not included in this analyses presented here). The pseudogene neutrality likelihood ratio test statistic for this human pseudogene as paired with the most closely related human and mouse genes in the non-redundant database was 210.43, an unreasonably improbable value for a χ^2_1 random variable. When the *est* database was searched, we found that it contained a mouse and human Calmodulin-like gene that grouped together with the Calmodulin 2 pseudogene in a maximum-likelihood tree to the exclusion of the previously assigned ortholog/paralog pair (data not shown). Using the new sequences from *est* database, PNLRT = 0.008 with new a branch length for the pseudogene close 0. We mention this results since this problem may be at high frequency in like datasets and will continue to pose a problem until the entire human and mouse

genomes have been completely sequenced, ordered, and annotated. One interesting result, though, is that using the pseudogene neutrality likelihood ratio test method in conjunction with data from a known phylogeny, it seems possible to find “missing” members of gene families that may not be in the gene databases or possibly even discovered.

The second hypothesis for the excess of pseudogenes that reject the pseudogene neutrality likelihood ratio test is that there is selection operating at the DNA or RNA level or that the mutational processes differ between the genomic regions of the functional gene and that of the pseudogene. The pseudogene neutrality likelihood ratio test tests the neutrality of a pseudogene at the *protein* level assuming a shared mutational process for the pseudogene in question and the ortholog from which it was derived. If selection is operating at the RNA or DNA level and not at the protein level or if there is variation in mutation rates among different parts of the genome, this will cause deviations from $\omega = 1$ along the pseudogene branch. This is particularly relevant to the discussion of mammalian pseudogene evolution since the mammalian genome is known to vary drastically in GC content from region to region. It is also known that functional genes tend to be non-randomly distributed with respect to GC content with a majority of genes found in GC-rich regions and a majority of the genome composed of AT-rich regions. If processed pseudogenes are randomly incorporated into the genome, the majority of pseudogenes will move to areas of lower GC content. The result of this is that pseudogenes derived from genes in GC-poor coding regions will land in genomic regions that are similar to the region in which their functional gene has evolved whereas genes from GC-rich regions will land in regions of lower GC content.

A prediction of this hypothesis is that the per silent site substitution rate in the pseudogene lineage should be elevated relative to the per replacement site substitution rate for

pseudogenes that move from regions of high GC content to regions of low GC content and vice versa. The reason for this is that we expect silent sites in *functional genes* to be most affected by local GC content since replacement sites are presumably under purifying selection. Processed pseudogenes that move from an area of high GC to low GC land with an “excess” of G or C encoded silent sites and are, therefore, more strongly subject to either mutation or selection pressure for GC content leading to an excess of silent substitutions vis a vis replacement substitutions. The same phenomena should hold for genes that move from low GC regions to high GC regions, but given the high abundance of AT rich regions and the fact that most genes are in GC rich regions, these events should occur infrequently.

As discussed above (fig. 5) we see that the prediction that $\omega < 1$ for pseudogenes that reject the pseudogene neutrality likelihood ratio test is upheld by the data. Likewise, a Wilcoxin signed rank test for equality of medians in substitution rates between silent and replacement sites suggests a higher rate of silent substitutions for pseudogenes derived from high GC regions ($Z = 4.333$; $p < 0.0001$) but not for genes derived from low GC region ($Z = -0.679$; $p \approx 0.5$). The replacement substitution rate is also found not to vary between GC regions ($Z = 0.407$; $p \approx 0.68$). Lastly, we find that overall pseudogenes derived from high GC regions tend to decrease in GC4 content ($Z = -3.561$; $p < 0.0005$) relative to their respective genes whereas pseudogenes derived from low GC regions tend to maintain the same level of GC4 content ($Z = -0.950$; $p \approx 0.3422$). These results tentatively support the hypothesis that the substitution rates in mammalian pseudogenes are strongly affected by whatever mechanism maintains variation in GC content in mammalian genomes and that this phenomena needs to be taken account of in estimating substitution rates for processed pseudogenes.

Implications of results for silent site evolution

Overall, we find that regardless of whether one considers functional genes from GC rich or GC poor regions, a significant number of genes have silent substitution rates that differ from the silent substitution rate in their respective pseudogenes (Table 1). This result is statistically significant whether one considers the proportion of genes that reject the silent site neutrality likelihood ratio test or whether one considers the pooled silent-site neutrality likelihood ratio test across genes. For genes in GC-rich regions this may be an artifact of an excess of silent site fixation events along the pseudogene branch in the gene tree due to selection or mutation biases in the novel genomic region. The same cannot be said for GC-poor genes, since we have shown that pseudogenes derived from GC-poor genes do not reject the pseudogene neutrality likelihood ratio test more than expected by chance alone and that substitution rates of silent and replacement mutations are roughly equal for pseudogenes in this class.

The major question of interest is how does the detection of non-neutral evolution at silent sites relate to the rate of substitution? Previous work has suggested that purifying selection may be rather strong on synonymous sites of murid and human genes (Ophir et al., 1999). One way to investigate this issue is to explore whether Model 5 is rejected because synonymous sites are evolving slowly when compared to the pseudogene or because they are evolving too fast.

In figure 5 we present histograms for the number of genes that reject or fail to reject the neutral silent site model for a given ds_f/ds_ψ range as calculated from Model 6. We find that the silent site neutrality likelihood ratio test rejects the neutral model in proportionally as many data sets because of large values of ds_f/ds_ψ as because of small values of ds_f/ds_ψ . This result holds regardless of whether one partitions the data based on the GC4 content of the functional gene. This suggests that the assumption of constrained evolution may, therefore, not be the whole of

the story for synonymous site evolution. Slightly advantageous mutations may also be driving an accelerated rate of evolution at certain silent sites.

To estimate the overall difference in the average rates of substitution between silent sites and pseudogene sites, we compare the ratio of the average rate of substitution per synonymous site in functional paralogs to the average rate of substitution per site in the pseudogenes ($\overline{ds_f}/\overline{ds_\psi}$). In Table 3 and Figure 6 we summarize estimates of and confidence intervals for $\overline{ds_f}/\overline{ds_\psi}$ based on estimates from model 6 (pseudogene neutrality model) and model 7 (free model) for gene-pseudogene pairs derived from GC poor regions and from model 7 for gene-pseudogene pairs derived from GC rich regions (we do not present the results from model 6 for the latter dataset, since this model has been rejected by the pooled pseudogene neutrality likelihood ratio test). The effect on the average rate of substitution seems to be moderate, on the order of a 30% reduction in the substitution rate of silent sites in functional genes relative to silent sites in pseudogenes. We also note that the estimates from GC rich regions for model 7 (0.62) differs slightly from the estimates for models 6 and 7 for GC poor regions, presumably due to a slight increase in the substitution rate in the pseudogene due to mutation or selection pressure. These estimates are all significantly different from unity (Table 3).

Once we have an estimate of the difference in substitution rate, we can estimate how the observed difference in substitution translates into selective differences within populations. From result (11) of Kimura (1962) one can easily show that for mutations with selection coefficient, s , and effective population size, N , the ratio of the rates of substitution at selected sites to neutral sites (assuming $|s|$ is small and that mutation rates are the same) is approximately given by

$$r \approx \frac{S}{1 - e^{-S}} \quad (5)$$

where $S = 4Ns$. If we assume that all mutations at silent sites have the same selective effect, that all pseudogene sites are neutral, and that the mutation rate is the same at pseudogenes and silent sites, then by setting $r = 0.70$ and solving (5) we find that the average selective effect is small, on the order of $S \approx -0.675$. This suggests that, at best, the average strength of selection in synonymous sites is in the range where genetic drift predominates in determining the course of evolution. On the other hand, if there is considerable variation within genes in selection intensity or if the mutation rates differ considerably between pseudogenes and functional genes at silent sites, the above result will underestimate the effect of selection. To estimate the effect of selection in these more complex scenarios, comparative data at both the within and between species level will be needed.

Acknowledgements

The authors would like to thank H. Akashi and D. Petrov for comments on earlier drafts of this manuscript. This work was supported by Howard Hughes Medical Institute award to CDB.

(NSF AWARD TO UCBERKLEY FOR COMPUTERS).

References

- Altschul, S.F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
- Andersson, S. G. E., and C. G. Kurland. 1990. Codon preferences in free-living microorganisms. *Microbiol. Rev.* 54: 198-210.
- Akashi, H. 1996. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. *Genetics* 144(3):1297-307.
- Akashi H. 1997. Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene*. 205(1-2):269-78.
- Akashi, H. and S. W. Schaeffer. 1997. Natural selection and the frequency distributions of "silent" DNA polymorphism in *Drosophila*. *Genetics* 146(1):295-307.
- Bensasson, D. D. A. Petrov, D. Zhang, D. L. Hartl, and G. M. Hewitt. 2001. Genomic Gigantism: DNA Loss Is Slow in Mountain Grasshoppers . *Mol Biol Evol* 18(2): 246-253.
- Bernardi, B. 2000. The compositional evolution of vertebrate genomes. *Gene* 259(1-2):31-43.
- Boguski, M. S., T. M. Lowe, and C. M. Tolstoshev. 1993. dbEST--database for "expressed sequence tags". *Nat. Genet.*(4):332-3.
- Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**:725-736.
- Gojobori, T., W. H. Li, and D. Graur. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18**:360-369.
- Graur, D., Y. Shuali, and W.-H. Li. 1989. Deletions in processed pseudogenes accumulate faster in murids than in humans. *J. Mol. Evol.* **28**:279-285.

- Hartl, D. L., M. Moriyama, and S. A. Sawyer. 1994. Selection intensity for codon bias. *Genetics* 138(1):227-34.
- Hartl, D. L., E. F. Boyd, C. D. Bustamante, and S. A. Sawyer. 2000. The glean machine: What can we learn from DNA sequence polymorphism. *Symp. Genomics & Proteomics* 4: 37-49.
- Kimura, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* 47: 713-719.
- Li, W. H., T. Gojobori, M. Nei. 1981. Pseudogenes as a paradigm of neutral evolution. *Nature* **292**(5820):237-9.
- Li, W. H., C. I. Wu, and C. C. Luo. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* **21**:58-71.
- McVean G. A., and J. Vieira. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. *Genetics* **157**:245-57.
- Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**:715-724.
- Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**:929-936.
- Ophhir, R., and D. Graur. 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* **205**: 191-202.
- Ophir, R., T. Itoh, D. Graur, and T. Gojobori. 1999. A Simple Method for Estimating the Intensity of Purifying Selection in Protein-Coding Genes. *Mol. Biol. Evol.* **16**:59-53.

- Petrov, D. A., E. Lozovskaya, and D. L. Hartl. 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**:346-9.
- Petrov, D. A., and D. L. Hartl. 1999. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc. Natl. Acad. Sci. USA* **96**:1475-1479.
- Petrov, D. A., T. A. Sangster, J. S. Johnston, D. L. Hartl, and K. L. Shaw. 2000. Evidence for DNA loss as a determinant of genome size. *Nature* **287**:1060-2.
- Sharp, P. M. and W.-H. Li. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**: 28-38.
- Smith, N. G. and Hurst L. D. The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate? *Genetics* **152**:661-73.
- Stuart, A., J. K. Ord, and S. Arnold. Kendall's Advanced Theory of Statistics Vol. 2A: Classical Inference and the Linear Model. Sixth Edition. Oxford University Press: New York.
- Xia, X. 2000. DAMBE: Data analysis in molecular biology and evolution. Department of Ecology and Biodiversity, University of Hong Kong.
- Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**:568-573.
- Yang, Z., and R. Nielsen. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* **46**:409-418.

Table 1: Results of LR tests and proportion of genes that reject null hypothesis for pooled data.

Dataset	Likelihood Ratio Test	Proportion of Test that Reject H_0 (95% Confidence Interval)	Pooled Test (P value)
Human Low Genomic GC	Pseudo	$1 / 34 = 0.0294$ (0.0017, 0.1231)	33.414 0.4962
	Silent site	$5 / 33 = 0.152$ (0.0571, 0.2978)	57.092 0.0057
Human High Genomic GC	Pseudo	$7 / 33 = 0.2121$ (0.0973, 0.3700)	71.754 0.0001
	Silent site	$7 / 26 = 0.2692$ (0.1258, 0.4560)	72.818 2.6×10^{-6}
Human Low GC4	Pseudo	$1 / 39 = 0.02564$ (0.0015, 0.1081)	39.37 0.4533
	Silent site	$9 / 38 = 0.2368$ (0.1216, 0.3864)	83.888 2.61×10^{-5}
Human High GC4	Pseudo	$8 / 35 = 0.2286$ (0.1117, 0.3838)	76.648 6.06×10^{-5}
	Silent site	$6 / 27 = 0.2222$ (0.0949, 0.3999)	75.338 1.88×10^{-6}
Rodent Low GC4	Pseudo	$1 / 22 = 0.04546$ (0.0027, 0.1852)	20.286 0.5651

	Silent site	3 / 21 = 0.1429 (0.0376, 0.3299)	56.798 3.85 X 10⁻⁵
Rodent High GC4	Pseudo	3 / 25 = 0.1200 (0.03129, 0.2836)	48.602 0.0032
	Silent site	7 / 22 = 0.3182 (0.1513, 0.5252)	53.63 0.0002

Table 2: Average ratio of substitution rates at synonymous sites in functional genes and in pseudogenes

Dataset (n)	$\overline{ds_f} / \overline{ds_\psi}$ 95% C.I.
Model 6 GC poor (61)	0.7071 (0.5801, 0.8523)
Model 7 GC poor (61)	0.6908 (0.4985, 0.9010)
Model 7 GC rich (60)	0.6232 (0.4518, 0.8482)

Figure 1. A graphical illustration of the model assumed in this paper. f is the codon sequence in the functional gene, ψ is the codon sequence in the pseudogene, o is the codon sequence in the outgroup, and t_f , t_ψ , t_o are the respective branch lengths of the lineages leading to these sequences.

Figure 2. Plots of the empirical CDF for the silent site neutrality likelihood ratio test (SSNLRT) and the pseudogene neutrality likelihood ratio test (PNLRT) versus the expected χ^2_1 distribution for pseudogenes derived from both GC rich and AT rich regions.

Figure 3. Problematic pseudogene phylogeny. A and B represent two different species, the filled and open letters represent duplicated genes, the unbroken line represents constrained (or adaptive) evolution, and the hatched line represents neutral evolution. We note that in this scenario if the “true” paralog and ortholog (\mathbb{A} and \mathbb{B}) of the pseudogene ($\Psi\mathbb{B}$) are not known, the resulting inference about the phylogeny will confound constrained or adaptive evolution with neutral evolution.

Figure 4. Relationship between ω in the putative pseudogene and frequency of rejecting the pseudogene neutrality likelihood ratio test. White bars are the number of genes that did not reject the test and shaded bars are the number of genes that did reject for a given range of ω .

Figure 5. Relationship between relative rate of substitution at synonymous sites and frequency of rejecting the silent site neutrality likelihood ratio test. White bars are the number of genes that

did not reject the test and shaded bars are the number of genes that did reject for a given range of the ratio ds_f/ds_ψ .

Figure 6. Distributions of $\overline{ds_f}/\overline{ds_\psi}$ estimated from 100,000 non-parametric bootstrap samples of estimates from model 6 (solid line) and model 7 (dotted line) from GC poor gene-pseudogene pairs and from model 7 (thick solid line) from GC rich pairs.

FIGURE 1

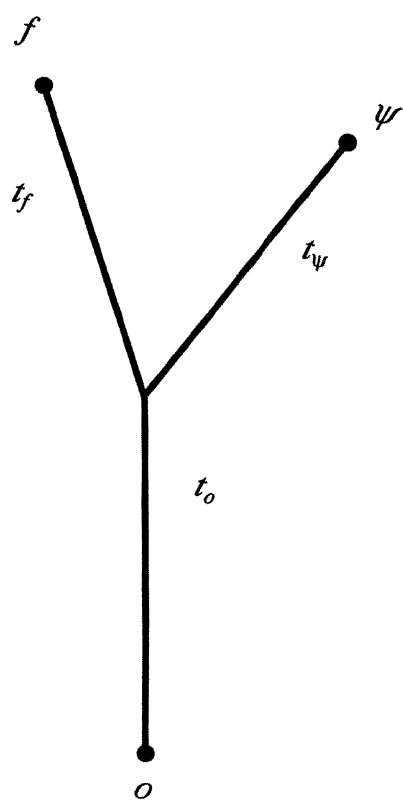


FIGURE 2

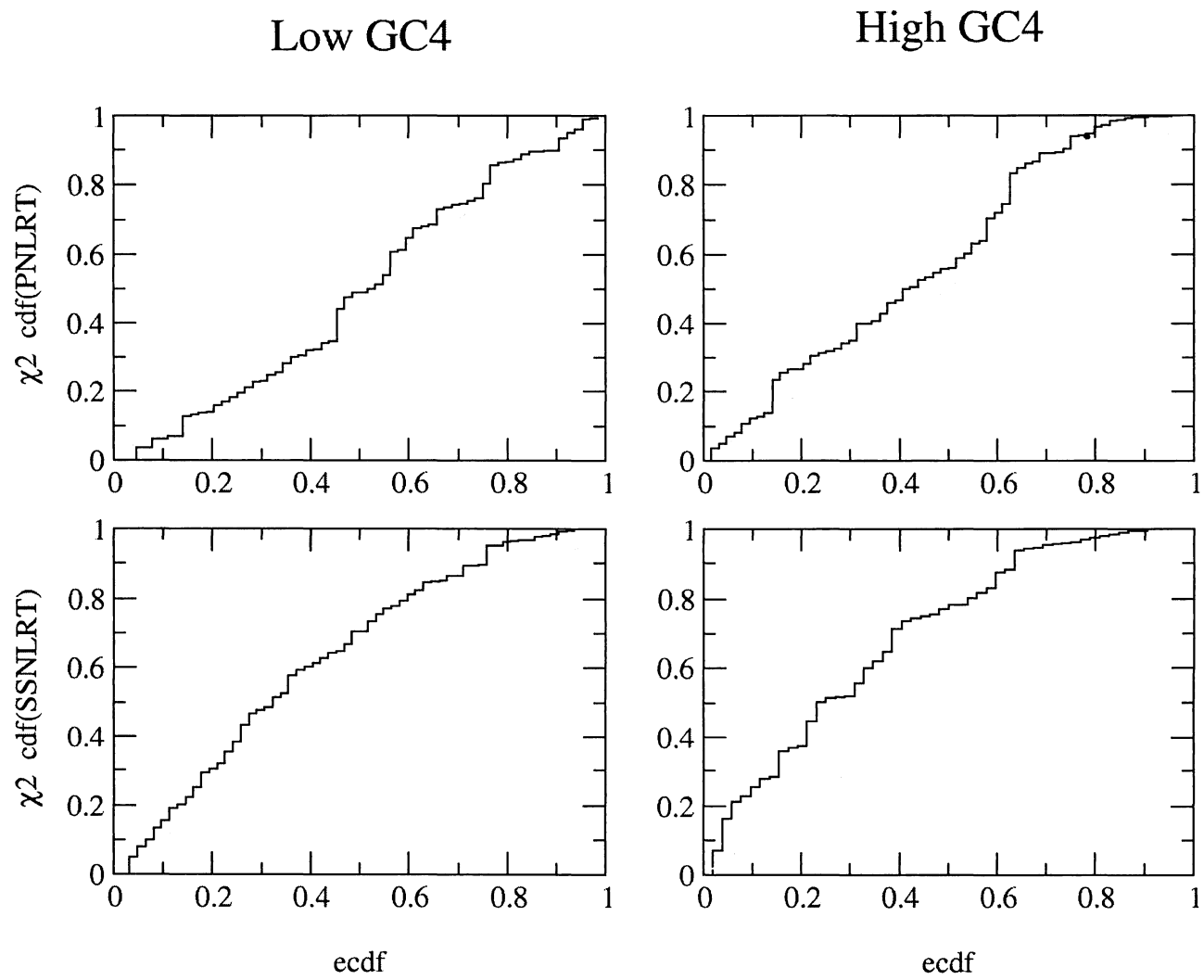


FIGURE 3

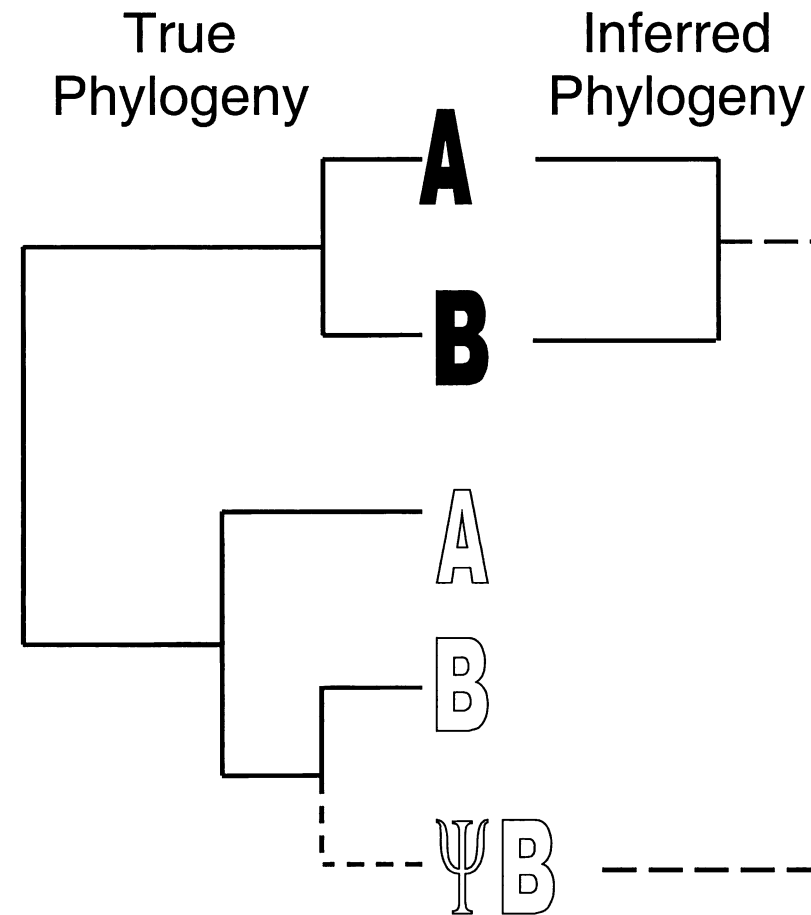


FIGURE 4

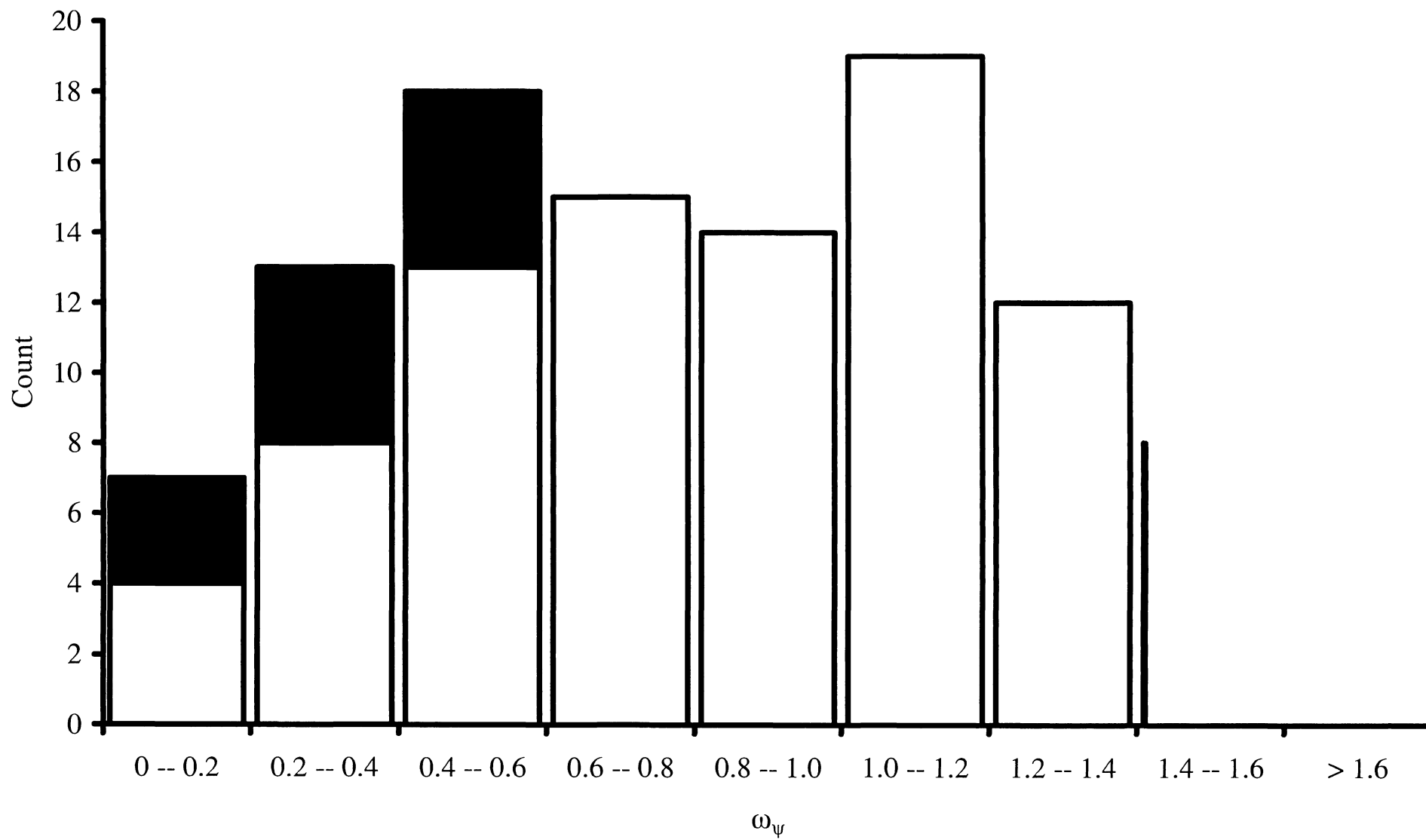


FIGURE 5

